°textA dissertation submitted in partial fulfillment of the requirements for the degree of

Certain observations concerning the effects of epistasis on complex traits and the evolution of genomes.

Maximilian Press

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Christine Queitsch, Chair

Elhanan Borenstein, Chair

First committee member

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Certain observations concerning the effects of epistasis on complex traits and the evolution of genomes.

Maximilian Press

Co-Chairs of the Supervisory Committee:
Associate Professor Christine Queitsch
Department of Genome Sciences

Associate Professor Elhanan Borenstein
Department of Genome Sciences

The informational content of genomes is usually interpreted mimetically, which is to say, with a one-to-one relationship between genotypes at certain genomic positions and phenotypic outcomes. While such interpretations have the virtue of simplicity, they are empirically unsuccessful in elucidating the working of biological systems and in allowing us to predict variation in many phenotypes. Many have called for such models to explicitly consider epistasis, which can roughly be defined as any consideration of interactions between genomic elements (a semiotic rather than mimetic interpretation of genomic information). In this thesis, I consider some of the consequences of the existence of such epistasis. In the first part of this thesis, I consider a particular case of a fast-evolving genetic element (short tandem repeat, or microsatellite) that shows widespread epistasis, and propose that such elements are likely to accumulate epistatic interactions by acting as mutational modifiers. I go on to show molecular mechanisms by which the element participates in epistasis, their phenotypic consequences, and make some observations on other short tandem repeats. In the second part of this thesis, I start with the assumption of epistasis between genes, and explore how this assumption can be used to understand the evolution of bacterial genome con-

tent. First, I take Hsp90, the known epistatic hub, and infer its coevolution with other genes through coordinated gains and losses across bacterial diversity. I further extend the underlying phylogenetic model to predict new "clients" of bacterial Hsp90, which have remained elusive when pursued through purely experimental approaches. Collaborators were able to validate certain of these predicted clients. Last, I attempt an analogy between prokaryotic genome evolution and the much better-understood field of protein evolution. I propose that, like protein evolution by substitution, genome evolution by horizontal acquisition of genes is substantially constrained by epistasis. I go on to infer the existence of such epistatic dependencies, where one gene in an ancestral genome promotes the acquisition of a second gene. A network of such dependencies shows a chronological structuring of gene acquisitions through prokaryotic evolution, suggesting universal assembly patterns by which genomes acquire functions. I go on to show that these dependencies are taxonomically universal (i.e. not restricted to particular phyla), and that they are sufficient to make reasonably good predictions about what genes a genome will gain in the future. This predictability of genome evolution by horizontal transfer confirms a major assertion of the protein evolutionists, that constraining epistasis leads to predictable evolutionary outcomes. Together, these observations indicate that the genetic architecture of traits and the content of genomes are shaped by the existence of networks of epistasis, reflecting the complex wiring of underlying biological functions.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

"I will ask you to mark again that rather typical feature of the development of our subject; how so much progress depends on the interplay of techniques, discoveries and new ideas, probably in that order of decreasing importance."

Sydney Brenner

That generation's dream, aviled
In the mud, in Monday's dirty light,

That's it, the only dream they knew,
Time in its final block, not time

To come, a wrangling of two dreams.
Here is the bread of time to come,

Here is its actual stone. The bread
Will be our bread, the stone will be

Our bed and we shall sleep by night.
We shall forget by day, except

The moments when we choose to play
the imagined pine, the imagined jay.

Wallace Stevens

DEDICATION

to my Sarah

Chapter 1

INTRODUCTION

1.1 The Purpose of This Sample Thesis

Chapter 2

BACKGROUND-DEPENDENT EFFECTS OF POLYGLUTAMINE VARIATION IN THE *ARABIDOPSIS THALIANA* GENE *ELF3*

Figures and tables prefixed with an 'S' can be found in Appendix A.

2.1 Summary

Tandem repeats (TRs) have extremely high mutation rates and are often considered to be neutrally evolving DNA. However, in coding regions, TR copy number mutations can significantly affect phenotype and may facilitate rapid adaptation to new environments. In several human genes, TR copy number mutations that expand polyglutamine (polyQ) tracts beyond a certain threshold cause incurable neurodegenerative diseases. PolyQ-containing proteins exist at a considerable frequency in eukaryotes, yet the phenotypic consequences of natural variation in polyQ tracts that are not associated with disease remain largely unknown. Here, we use Arabidopsis thaliana to dissect the phenotypic consequences of natural variation in the polyQ tract encoded by *EARLY FLOWERING 3* (*ELF3*), a key developmental gene. Changing ELF3 polyQ tract length affected complex ELF3-dependent phenotypes in a striking and non-linear

manner. Some natural ELF3 polyQ variants phenocopied elf3-loss-function mutants in a common reference background, although they are functional in their native genetic backgrounds. To test the existence of background-specific modifiers, we compared the phenotypic effects of ELF3 polyQ variants between two divergent backgrounds, Col and Ws, and found dramatic differences. Our data support a model in which variable polyQ tracts drive adaptation to internal genetic environments.

2.2 Introduction

In coding regions, tandem repeat (TR) copy number variation can have profound phenotypic effects [3]. For example, TR copy number mutations that expand polyglutamine (polyQ) tracts past a threshold number of glutamines can cause incurable neurodegenerative diseases such as Huntington's disease and Spinocerebellar Ataxias [2, 4]. PolyQ tract length correlates with onset and severity of polyQ expansion disorders, but for intermediate polyQ tracts this correlation is far weaker (4-8), suggesting the existence of genetic and environmental modifiers (9-12). Despite their potential for pathogenicity, variable polyQ tracts occur frequently in eukaryotic proteins, many of them functioning in development and transcription (1, 13-15). Model organism studies have suggested that coding TRs are an important source of quantitative genetic variation that facilitates evolutionary adaptation (1, 16-19). For example, TR copy number variation in the yeast gene FLO1 correlates linearly with flocculation (20), a phenotype that is important for stress survival (17). As polyQ tracts often mediate protein interactions (2, 3, 21), polyQ-encoding TR copy number mutations could produce large and possibly adaptive phenotypic shifts. To determine the phenotypic impact of naturally occurring polyQ variation (18, 22, 23) in a genetically tractable model, we focused on the gene ELF3, which encodes a polyQ tract that is highly variable across divergent Arabidopsis thaliana strains (accessions) (19, 24). ELF3 is a core component of the circadian clock and a potent repressor of flowering, and is considered a "hub protein" for its many interactions with various proteins (24-31). Consequently, elf3 loss-of-function

mutants show pleiotropic phenotypes: they flower early, show poor circadian function, and grow long embryonic stems (hypocotyls) in light (25-27, 29, 30, 32). Single nucleotide polymorphisms (SNPs) in ELF3 affect shade avoidance, a fitness-relevant plant trait (24, 33). ELF3 polyQ variation has been suggested to correlate with two parameters of the circadian clock, period and phase (19). The ELF3 polyQ tract may mediate ELF3 membership in protein complexes, though thus far no ELF3-binding protein is known to bind it (26, 28-30). We discovered that altering polyQ tract length has dramatic effects on ELF3-dependent phenotypes and that these effects are dependent on genetic background.

2.3 Methods

2.3.1 Plant Materials and Growth Conditions.

The 181 Arabidopsis thaliana accessions are as previously described (1). The loss-of-function EARLY FLOWERING 3 (elf3) mutants are: (i) *elf3-4*, containing a CCR2::LUC transgene (ecotype Ws) (2, 3) and 2) *elf3-200*, the GABI750E02 T-DNA insertion mutant (ecotype Col-0) (4). For hypocotyl experiments, seeds were sterilized with Ethanol and plated onto 1x Murashige and Skoog (MS) basal salt medium supplemented with 1x MS vitamins, 1% sucrose, 0.05% Mes (wt/vol), and 0.24% (wt/vol) phytagel. After stratification in the dark at 4° C for 3 d, plates were transferred to an incubator (Conviron) that was set to either short day (SD) (8L:16D at 20° C) or long day (LD) (16L:8D at 22° C : 20° C), with light supplied at 100μmol $* $ m² $*$ s^{-1} by cool-white fluorescent bulbs. For growth on soil, seeds were stratified at 4° C for 3 d, and then grown in Sunshine #4 soil under cool-white fluorescent light at either LD or SD at 20 °C. Seedlings used for RNA extractions were grown on soil under LD conditions and harvested on day 10. Samples for ELF3 expression measurements were collected at Zeitgeber time (ZT) 20. Samples for Phytochrome- interacting Factor 5 (PIF5) expression measurements were collected a ZT 8. Samples for and Pseudoresponse regulator 9 (PRR9) expression measurements

were collected at ZT 0, 5, and 8.

2.3.2   *Generation of ELF3 Transgenic Plants.*

To generate A. thaliana transgenics carrying different ELF3 tandem repeat (TR) alleles, the cDNA clone RAFL09-28-E05 (RIKEN BRC) (5, 6), containing the ELF3 coding region and 3' UTR (Col-0 accession) was used. This cDNA clone lacks the small 5' intron. Two re- striction sites, NarI and NcoI, were inserted into the ELF3 coding sequence using the QuikChange Site-Directed Muta- genesis kit (Stratagene) (primer information in Table S5). The polyglutamine (polyQ)-encoding region was amplified from ac- cessions containing selected TR copy number alleles (primer information in Table S5, TR allele information in Table S1). These PCR products were digested with NarI/NcoI and ligated into the previously mutagenized ELF3 coding region. An artifi- cial allele lacking the TR was generated by site-directed muta- genesis (primer informa- tion in Table S5). Mutated plasmids and all ligation products were sequenced to ensure accuracy. The ELF3 alleles were cloned into pENTR1A (Invitrogen). A 2-kbp NotI fragment containing the ELF3 promoter was inserted up- stream of each ELF3 coding sequence. The fragments containing the ELF3 promoter, ELF3 coding sequence, and the ELF3 3' UTR were recombined using Gateway LR Clonase II (In- vitrogen) into a modified pB7WG2 (7), which lacks the CaMV- 35S promoter. The region encoding the polyQ tract of each construct was sequenced to ensure accurate TR copy number. The plasmids were used to transform Agrobacterium tumefaciens GV3101. Subsequently, Arabidopsis elf3 mutants were trans- formed by the flower dip method (8). Transfor- mants were se- lected on Basta (Liberty herbicide; Bayer Crop Science) and propagated for three to four generations. The accuracy of the transgenes was confirmed by PCR (primer information in Table S5). All Ws phenotypic assays were performed in homozy- gous transgenic plants with expression levels between 0.8- and 4.5- times the respective ELF3 wild-type (Fig. S1C); for Col lines, transgene expression levels were between 0.3- and 4.3-times the respective ELF3 wild-type (Fig. S1D). Analyzed plant lines are in Ta-

bles S2–S4.

2.3.3 *RNA Extractions and Real-Time PCR.*

Total RNA was extracted from 30-mg frozen tissue using the SV Total RNA Isolation System (Promega). Subsequently, 2 μg of RNA were subjected to DNase treatment using Ambion Turbo DNA-free Kit (Applied Biosystems). RNA integrity and purity were checked with an Agilent Bioanalyzer using the RNA 6000 Nano Kit (Agilent Technologies). For cDNA synthesis, 200 ng of DNase-treated RNA was reverse-transcribed using the Transcriptor First Strand cDNA Synthesis Kit (Roche) and oligo dT primers. Transcript abundance was determined by real-time quantitative PCR using the Light-Cycler 480 system (Roche), with LightCycler 480 SYBR Green I Master (Roche) and the following PCR conditions: 5 min at 95 °C, followed by 35 cycles of 15 s at 95 °C, 20 s at 55 °C, and 20 s at 72 °C. To ensure that PCR products were unique, a melting- curve analysis was performed after the amplification. UBC21 expression (At5g25760) was used as a reference. All quantitative RT- PCR primers were designed with the LightCycler Probe Design Software (Roche). Sequences for real-time PCR primers are shown in Table S6. Relative quantification was determined with the $\Delta\Delta C_T$ Method (9). Error was calculated as previously described (10).

2.3.4 *Thermal Asymmetric Interlaced PCR.*

High-efficiency thermal asymmetric interlaced (TAIL)-PCR was performed as previously described (11) to obtain the flanking sequence of the construct in- tegration site (left border). Briefly, a preamplification step was performed with primers LAD and LB-0a (Table S7), followed by primary TAIL-PCR with primers AC1 (11) and LB-1a (Table S7), and 1 μL of a 1/40 dilution of the preamplification product as a template. A secondary TAIL-PCR with primers AC2 (11) and LB-2a (Table S7) was performed with 1 μL of a 1/10 dilution of the primary TAIL-PCR product. Next, 3-kbp products were extracted from agarose gels and subsequently Sanger-sequenced. Only sequences

containing the T-DNA left border were considered.

2.3.5 *Developmental Phenotype Assays.*

For measurements of hypocotyl length, seedlings were grown on vertical plates for 15 d in a pseudorandomized design under either SD or LD conditions (12). Hypocotyl length was measured with ImageJ on digital images (http://rsbweb.nih.gov/ij/). For measurement of flowering time, seeds were planted in sheet pots (36 pots per tray) in a randomized design and trays were rotated daily. Flowering time was re- corded as the day when the inflorescence reached 1 cm in height. Rosette leaf number was determined on the same day. Petiole- length/leaf-length (PL/LL) ratio for leaf four was determined on day 45. Least-square means for all traits were derived from a linear regression analysis for each trait separately. ELF3 TR copy number was modeled as a nominal variable and independent transgenic lines carrying the same ELF3 TR allele were analyzed together. We tested for significant phenotypic differences con- ferred by the different ELF3 TR alleles by using Tukey-HSD tests with $\alpha = 0.05$ that accommodate nonnormal data.

2.3.6 *Luciferase Imaging and Period Analysis.*

Luciferase assays were performed with lines containing the CCR2::LUC reporter. Seeds were surface sterilized with a 70% (vol/vol) ethanol wash followed by a second wash with 33% (vol/vol) Klorix with Triton X-100, and then rinsed twice with sterile water. Seeds were plated on MS3 medium [pH 5.7, 3% (wt/vol) sucrose, 1.5% (wt/vol) PhytoAgar, and 15 μg/mL hygromycin B]. They were subsequently stratified for 4 d at 4 °C in the dark and entrained under 12-h light:12-h dark cycles under white fluorescent light ($\sim 10\mu$mol $* $ m^{-2} $* $ s^1) at 22 °C. On the sixth day, a minimum of 24 seedlings per line was transferred to 96-well TopCount (Perkin-Elmer) plates containing 200 mg MS3 agar. We added 5 mM Luciferin in 0.01% Triton X-100 and entrained seedlings for another cycle before luminescence was detected using a Packard/Perkin-Elmer TopCount Scintillation and Luminescence Counter. Red and blue light-emitting diodes (

100μmol $* \mathrm{m}^{-2} * \mathrm{s}^{-1}$) were used as a light source during this analysis. During the first 24 h of luminescence detection, plants were grown in 12-h light:12-h dark and then released under constant light conditions to measure the free- running period. Each individual was measured approximately every 30 min for a minimum of 5 d. Luminescence levels were quantified and analyzed as previously described (2, 3) using the macro suites TopTempII and Biological Rhythms Analysis Software System (13). Period length and relative amplitude error (RAE) were estimated using fast Fourier transform nonlinear least squares (14). Period values scored with RAE values below 0.4 were considered robustly rhythmic (15).

2.3.7 *Principal Component Analysis.*

We clustered our phenotypic data using principal component analysis (PCA) to find patterns corresponding to genotypes. We excluded the phenotype of rosette leaf number in SD, for which data were missing for several alleles. The phenotypes included in the analysis are: Days to flowering in SD and LD conditions, hypocotyl length under SD and LD PL/LL for the fourth leaf in SD, and rosette leaf number in LD. For analyses involving Col lines, the SD PL/LL ratio phenotype was omitted because of lack of data, and PCA was thus based on the remaining five phenotypic variables. For each phenotype in each genetic background (either Ws or Col-0), we calculated the mean phenotype of the independently generated lines for each *ELF3-TR* allele, giving us a 28 x 6 matrix of mean phenotypes for the 28 genotypes for each of six phenotypic variables. Within each background, we ranked the genotypes for each phenotype. Ranks were transformed into a standard normal distribution based on their percentile, using the R function qnorm. Using this transformed dataset, we performed PCA using the R function *prcomp* (R Foundation for Statistical Computing, http://www.r-project.org/, 2011). We performed PCA for each background separately, and then for both backgrounds together. Rank-normalization was necessary to compare (i) phenotypes measured on different scales and (ii) Ws- and Col-derived plants, between which backgrounds ab-

solute phenotypic differences exist. Consequently, the rank-normalization increases stability of our estimates, as our dataset is relatively small and PCA☐ as assumptions of normality were not met by our raw dataset. PCA on raw values scaled to a standard normal distribution gave similar results. Biplots were generated with the R *biplot* function on *prcomp* function output.

2.4   ELF3-TR variation affects ELF3-dependent phenotypes.

Among 181 natural A. thaliana accessions, the *ELF3-TR* encoded between 7 and 29Q (Table S1, Fig. S1a). For comparison, polyQ expansions over 20Q are associated with disease in the context of the SCA6 gene, though most other disease-associated polyQ expansions are longer (2, 19, 24). The most frequent *ELF3-TR* encoded 16Q, whereas the shortest TR (7Q) was found in the reference strain Col-0. We set out to test whether naturally occurring *ELF3-TR* alleles affect ELF3-dependent phenotypes and whether they do so in a linear manner as suggested by association studies (19) and found for coding TR variation in other genes (16, 20). We generated expression-matched transgenic lines for most natural *ELF3-TR* alleles in the loss-of-function *elf3-4* mutant (Ws background, Table S2a, Fig. S1c) (32) and measured their flowering time and circadian clock-related phenotypes (Figs. 1, S2a-g). *ELF3-TR* variation significantly affected ELF3-dependent phenotypes, but there was no evidence of a linear relationship. The different *ELF3-TR* alleles resulted in phenotypes ranging from nearly full complementation of *elf3-4* to nearly phenocopying the loss-of-function mutant. We used principal components analysis to describe the complex effects of *ELF3-TR* alleles on all tested ELF3-dependent phenotypes (PCA, Figs. 1a, S2h-j). Principal component 1 (PC1) corresponds to general functionality of ELF3 in all measured phenotypes, with wild-type Ws and mutant *elf3-4* defining the extremes. Separation along PC1 is driven by the tendency of plants with functional ELF3 to show short hypocotyls, late flowering, increased rosette leaf number, and short petioles (Figs. 1b-d, S2). The endogenous ELF3-16Q allele complemented both the early-flowering and long-hypocotyl phenotypes of

elf3-4 (Figs. 1b-d, S2). In contrast, both the long ELF3-23Q and the short ELF3-7Q allele (endogenous TR alleles in Br-0/Bur-0 and Col-0, respectively) behaved similarly to the *elf3-4* loss-of-function allele (Figs. 1b-d, S2), although they are functional in their native backgrounds. Neither Col-0 nor Br-0 and Bur-0 show the phenotypic characteristics of elf3-mutants (early flowering (34), long hypocotyls (35) and long petioles (36)), suggesting that *ELF3-TR* alleles may interact with background-specific modifiers. ELF3-0Q, an artificial ELF3 allele lacking the TR, partially complemented *elf3-4* (Figs. 1a, S2). Hence, the polyQ-encoding TR is not necessary for all ELF3 function, but changes in TR copy number are sufficient to enhance or ablate ELF3 function. PC2 separated ELF3-20Q and ELF3-11Q, which behaved as hypomorphs in certain phenotypes but not others (Fig. 1a). For example, ELF3-20Q plants had significantly longer hypocotyls than wild-type and its petioles phenocopied the extremely long petioles of the *elf3-4* mutant (Fig. 1c-e), but they did not differ from wild-type in flowering time (days to flower, Fig. 1b). The existence of both general and specific hypomorphs suggests that polyQ variation affects the multiple ELF3 functions separately. As part of a protein complex, ELF3 affects expression of Phytochrome-interacting Factor 5 (PIF5) and Pseudo-response regulator 9 (PRR9) (28, 37, 38). PIF5 and PRR9 expression were strongly affected by ELF3 polyQ variation (Fig. S3). ELF3-16Q phenocopied wild-type PRR9 and PIF5 expression, and the hypomorphic ELF3-23Q phenocopied *elf3-4* (28, 37, 38), mirroring their developmental phenotypes. Consistent with their divergence along PC2 (Fig. 1a), ELF3-11Q and ELF3-20Q differed in their effect on PRR9 expression, but not on PIF5 expression (Fig. S3a, b), demonstrating that ELF3 polyQ variation differentially affects the regulation of downstream genes.

2.5 ELF3-TR-*variation modulates the precision of the circadian clock*

To directly assess the role of ELF3 polyQ variation in the circadian clock, we used the CCR2::LUC reporter system (25, 39). We observed little difference in circadian period among wild-type Ws and tested ELF3-TR alleles (Fig. S4a), contradicting a pre-

viously observed association of TR copy number with period in natural accessions (19). However, we found that the relative amplitude error (RAE) of oscillation varies substantially across ELF3-TR genotypes (Figs. 2a, S4b). RAE measures the precision of a circadian period (40): high RAE values (> 0.4) indicate poor oscillation and clock dysfunction (41). The endogenous Ws ELF3-16Q nearly complemented the elf3-4 RAE defect, whereas the TR alleles ELF3-7Q, ELF3-10Q, and ELF3-23Q showed higher RAE, approaching arrhythmic *elf3-4* levels (Fig. 2a, b), consistent with their hypomorphic performance in other ELF3 traits (close to *elf3-4* in PC1, Fig. 1a). Together, these results suggest that ELF3 polyQ tract length is a critical determinant of circadian clock precision, but not period length, in A. thaliana.

2.6 ELF3-TR variation interacts with genetic background.

To test our hypothesis that *ELF3-TR* variation interacts with genetic background, we regenerated all *ELF3-TR* transgenic lines in the *elf3-200* loss-of-function mutant with matched transgene expression (Col background, Table S2c, Fig. S1d) (42). We used PCA to compare *ELF3-TR* effects between Ws and Col backgrounds (Figs. 3a, S5). The Col-specific ELF3-7Q allele complemented *elf3-200* in some traits such as flowering time (in short days, SD) and hypocotyl length (in long days, LD), but not others (Figs. 3a, b, S5, S6). This result may be due to the absence of the small 5' intron from the ELF3 construct used in this study. However, there was still a dramatic spread of phenotypes: all longer *ELF3-TR* alleles (>20 Qs) nearly complemented *elf3-200*, delaying flowering and shortening hypocotyls, whereas few of the shorter alleles did (Figs. 3, S5, S6). Results were similar when the Col data were analyzed alone (Fig. S6). Thus, in contrast to our results in the Ws background, *ELF3-TR*s appeared to show a threshold effect for TR copy number in the Col background. We speculate that the intensive laboratory propagation of the Col-0 accession may have altered selection on the *ELF3-TR*, resulting in an extremely short "hypomorphic" allele, whereas under natural conditions a longer TR might be more functional. Comparing TR allele effects between the two

backgrounds revealed striking differences. For example, the ELF3-23Q allele was a general hypomorph in the Ws background (*elf3-4*), whereas it produced highly functional ELF3 in the Col background (*elf3-200*, Fig. 3). In turn, the ELF3-16Q allele produced highly functional ELF3 in the Ws background (*elf3-4*), but was a general hypomorph in the Col background (*elf3-200*). The consistent performance of the artificial ELF3-0Q allele across backgrounds suggests that the background effect is TR-dependent (Figs. 3a, S5). Collectively, our results support that *ELF3-TR* alleles interact with background-specific modifiers.

2.7 *Col* ELF3 *allele is not haploinsufficient in Col x Ws hybrids.*

To address whether Ws and Col-specific background effects are sufficient for altered hybrid phenotypes, we generated F1 populations between wild-type and elf3 null plants in the Ws and Col backgrounds and measured ELF3 function by assessing hypocotyl length. Ws x Col F1 hybrids resembled their wild-type parents (Fig. 4). F1 hybrids containing both loss-of-function alleles had significantly longer hypocotyls than either parent (Fig. 4). Both ELF3 alleles were haplosufficient in F1 crosses within their native backgrounds, as expected for recessive mutants (Fig. 4). In stark contrast, we observe that ELF3-Col, but not ELF3-Ws, phenocopied the extreme hypocotyl length of the double loss-of-function mutant (Fig. 4). Consistent with the results from our transgenic lines, our F1 hybrid data suggest that full ELF3 function depends on a permissive genetic background.

Unfortunately, propagation of these F_1 hybrids to the F_2 generation and subsequent Col x Ws crosses revealed that the data in Figure 4 do not generalize to other crosses, and probably represent a spontaneous mutation in the Col background leading to ELF3 inactivation. In the face of such equivocal evidence, we suggest that more intensive genetic or biochemical experiments will be necessary to determine the relevant background modifiers of *ELF3-TR* variation. For two such approaches, refer to Chapter 5.

2.8   Discussion

Our results demonstrate that natural ELF3 polyQ variation that is not associated with disease has dramatic phenotypic consequences, and that these consequences depend on genetic background. For ELF3, in at least the Ws background, the relationship between TR copy number and phenotype does not follow a linear or threshold pattern as observed for other coding TR and polyQ disorders (1, 2, 16, 17, 20). Studies correlating TR variation with phenotype often apply linear models, treating TR copy number as a quantitative variable (19, 22, 23). Our data show that this approach is not appropriate for all TRs. Instead, *ELF3-TR* alleles seem "matched" to specific genetic backgrounds, in which they are functional, whereas they are incompatible with other backgrounds. The haploinsufficiency of the elf3-col allele in Ws x Col hybrids supports this interpretation. In contrast, the ELF3-Ws allele is haplosufficient in hybrids, indicating that the ELF3-Col x Ws incompatibility is asymmetric. This observation agrees with Orr's assertion that incompatibility between recently diverged populations is usually asymmetrical, because it tends to arise from the derived allele (i.e. ELF3-Col) (43). Variable TRs, and the *ELF3-TR* in particular, have been previously suggested as agents of adaptation to new external environments (1, 16, 17, 20, 24, 44). Our results suggest that polyQ-encoding TRs are also agents of coadaptation within genomes. We speculate that the observed background effects arise from background-specific polymorphisms in genes encoding physically interacting proteins (26, 28-30). TRs have a far higher mutation rate than non-repeated regions (10^{-4} per site per generation for TR vs. 10^{-8} for SNPs) (45, 46) and, as we show, their expansion or contraction can have dramatic phenotypic impact. ELF3's partner proteins may have acquired compensatory mutations to accommodate new *ELF3-TR* variants and vice versa. Alternative explanations for the background effects are compensatory mutations in ELF3 (intragenic suppressors), or ELF3 interactions that are unique to a given background. Intragenic variation and protein modification can play an important role in polyQ-mediated phenotypes (47,

48). At least for the ELF3-Col allele, however, our F1 data are not consistent with intragenic suppressors. Consistent with polyQ-mediated background effects, in at least one case, a modifier mutation has been shown to delay onset of Huntington's disease (11). Hypothetically, population genetic approaches could identify incompatible alleles that may contribute to variable disease onset in patients with polyQ expansions and to ELF3-dependent background effects in A. thaliana. However, the great diversity of TR alleles compared to SNP alleles and the small number of individuals carrying specific TR alleles render a population genetics approach infeasible. Extensive genetic mapping or other experimental approaches will be needed to identify the determinants of *ELF3-TR* dependent background effects. As TRs are rapidly evolving, we speculate that polyQ-mediated incompatibilities and the resulting fitness loss in hybrids and their offspring may contribute to disruption of gene flow between closely related populations. This speciation mechanism would be of particular importance for organisms with many polyQ-encoding TRs, thousands of offspring, and an inbreeding life style. Even in humans, however, about 1% of proteins contain polyQ tracts (13, 14, 45). Our results identify TR copy number variation, and in particular polyQ variation, as a phenotypically important class of genetic variation that warrants genome-wide assessment in model organisms, crops, and humans alike.

Chapter 3

THE CONSERVED *PFT1* TANDEM REPEAT IS CRUCIAL FOR PROPER FLOWERING IN *ARABIDOPSIS THALIANA*

A version of this chapter was published under the following reference:

Pauline Rival, Maximilian O. Press, Jacob Bale, Tanya Grancharova, Soledad F. Undurraga, and Christine Queitsch. The Conserved PFT1 Tandem Repeat is Crucial for Proper Flowering in *Arabidopsis thaliana. Genetics*, 198(2):747-754, August 2014.

Figures and tables prefixed with an 'S' can be found in Appendix B.

Chapter 4

SHORT TANDEM REPEATS AND QUANTITATIVE GENETICS

Portions of this chapter were published under the following references:

- Maximilian O. Press, Keisha D. Carlson, and Christine Queitsch.

  The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics*, 30(11):504-512, August 2014.

- Keisha D. Carlson, Peter H. Sudmant, Maximilian O. Press, Evan E. Eichler, Jay Shendure, and Christine Queitsch. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Research*, 25(5):750-761, May 2015.

Figures and tables prefixed with an 'S' can be found in Appendix C.

Chapter 5

THE VARIABLE ELF3 POLYGLUTAMINE HUBS AN EPISTATIC NETWORK

Figures and tables prefixed with an 'S' can be found in Appendix D.

Chapter 6

ELF3 POLYGLUTAMINE VARIATION REVEALS A PIF4-INDEPENDENT ROLE IN THERMORESPONSIVE FLOWERING

Figures and tables prefixed with an 'S' can be found in Appendix E.

Chapter 7

GENOME-SCALE CO-EVOLUTIONARY INFERENCE IDENTIFIES FUNCTIONS AND CLIENTS OF BACTERIAL HSP90

A version of this chapter was published under the following reference:

Maximilian O. Press, Hui Li, Nicole Creanza, Guenter Kramer, Christine Queitsch, Victor Sourjik, and Elhanan Borenstein. Genome-scale co-evolutionary inference identifies functions and clients of bacterial Hsp90. *PLoS Genetics*, 9(7):e1003631, 2013.

Figures and tables prefixed with an 'S' can be found in Appendix F.

<div align="center">Chapter 8</div>

EVOLUTIONARY ASSEMBLY PATTERNS OF PROKARYOTIC GENOMES

A version of this chapter is under review for publication, and is available at http://biorxiv.org/content/early

Figures and tables prefixed with an 'S' can be found in Appendix G.

Chapter 9

THE THESIS UNFORMATTED

This chapter describes the uwthesis class (`uwthesis.cls`, version dated 2011/06/27) in detail and shows how it was used to format the thesis. A working knowledge of Lamport's LATEX manual[**?**] is assumed.

9.1 *The Control File*

The source to this sample thesis is contained in a single file only because ease of distribution was a concern. You should not do this. Your task will be much easier if you break your thesis into several files: a file for the preliminary pages, a file for each chapter, one for the glossary, and one for each appendix. Then use a control file to tie them all together. This way you can edit and format parts of your thesis much more efficiently.

Figure 9.1 shows a control file that might have produced this thesis. It sets the document style, with options and parameters, and formats the various parts of the thesis—but contains no text of its own.

The first section, from the `\documentclass` to the `\begin\{document\}`, defines the document class and options. This thesis has specified two-sided formatting, which is now allowed by the Graduate School. Two sided printing is now actually LATEX's default. If you want one sided printing you must specify `oneside`. This sample also specified a font size of 11 points. Possible font size options are: `10pt`, `11pt`, and `12pt`. Default is 12 points, which is the preference of the Graduate School. If you choose a smaller size be sure to check with the Graduate School for acceptability. The smaller fonts can produce very small sub and superscripts.

Include most additional formatting packages with `\usepackage`, as describe by Lamport[**?**].

Figure 9.1: A thesis control file (`thesis.tex`). This file is the input to LaTeX that will produce a thesis. It contains no text, only commands which direct the formatting of the thesis. This is also an example of a 'facing page' caption. It is guaranteed to appear on a lefthand page, facing the figure contents on the right. See the text.

```
% LaTeX thesis control file

\documentclass[11pt,twoside]{uwthesis}

\begin{document}

% preliminary pages
%
\prelimpages
\include{prelim}

% text pages
%
\textpages
\include{chap1}
\include{chap2}
\include{chap3}
\include{chap4}

% bibliography
%
\bibliographystyle{plain}
\bibliography{all}

% appendices
%
\appendix
\include{appxa}
\include{appxb}

\include{vita}
\end{document}
```

The one exception to this rule is the `natbib` package. Include it with the `natbib` document option.

Use the `\includeonly` command to format only a part of your thesis. See Lamport[**?**, sec. 4.4] for usage and limitations.

9.2   The Text Pages

A chapter is a major division of the thesis. Each chapter begins on a new page and has a Table of Contents entry.

9.2.1   Chapters, Sections, Subsections, and Appendices

Within the chapter title use a `\\` control sequence to separate lines in the printed title (recall Figure **??**.). The `\\` does not affect the Table of Contents entry.

Format appendices just like chapters. The control sequence `\appendix` instructs LaTeX to begin using the term 'Appendix' rather than 'Chapter'.

Sections and subsections of a chapter are specified by `\section` and `\subsection`, respectively. In this thesis chapter and section titles are written to the table of contents. Consult Lamport[**?**, pg. 176] to see which subdivisions of the thesis can be written to the table of contents. The `\\` control sequence is not permitted in section and subsection titles.

9.2.2   Footnotes

Footnotes format as described in the LaTeX book. You can also ask for end-of-chapter or end-of-thesis notes. The thesis class will automatically set these up if you ask for the document class option `chapternotes` or `endnotes`.

If selected, chapternotes will print automatically. If you choose endnotes however you must explicitly indicate when to print the notes with the command `\printendnotes`. See the style guide for suitable endnote placement.

9.2.3 Figures and Tables

Standard LaTeX figures and tables, see Lamport[**?**, sec. C.9], normally provide the most convenient means to position the figure. Full page floats and facing captions are exceptions to this rule.

If you want a figure or table to occupy a full page enclose the contents in a `fullpage` environment. See figures 9.2.

Facing page captions are described in the Style Manual[**?**]. They have different meanings depending on whether you are using the one-side or two-side thesis style.

If you are using the two-side style, facing captions are full page captions for full page figures or tables and must face the illustration to which they refer. You must explicitly format both pages. The caption part must appear on an even page (left side) and the figure or table must come on the following odd page (right side). Enclose the float contents for the caption in a `leftfullpage` environment, and enclose the float contents for the figure or table in a `fullpage` environment. Figure 9.1, for example, required a full page so its caption (on a facing caption page) would have been formatted as shown in figure 9.2a. The first page (left side) contains the caption. The second page (right side) could be left blank. A picture or graph might be pasted onto this space.

If instead you are using the one-side style, facing caption pages are still captions for full page figures or tables that appear on the left-hand page (facing the illustration on the right-hand page). However, the page number and binding offset are reversed from their normal positions. Format these captions by enclosing the float contents in a `leftfullpage` environment. Because you are printing on only one side of each sheet, you must manually turn over this caption sheet. You then have the choice of inserting a preprinted illustration or formatting one to print with the thesis. In either case no page number should appear on the illustration page, nor should the page number increment. Enclose your figure's text in an `xtrafullpage` environment, which will cause the page numbers to come out right. You can, of course, leave out the illustration and insert a

```
\begin{figure}[p]% the left side caption
  \begin{leftfullpage}
    \caption{ . . . }
  \end{leftfullpage}
\end{figure}
\begin{figure}[p]% the right side space
  \begin{fullpage}
    . . .
      ( note.. no caption here )
  \end{fullpage}
\end{figure}
```

Figure 9.2a: This text would create a double page figure in the two-side style.

preprinted copy later. Figure 9.2b shows how to format a facing caption page in the one-side style. Note that, in this case, the illustration was also printed.

In the two-side style the xtrafullpage environment acts just like the fullpage environment. It does not produce a numberless page.

9.2.4 *Horizontal Figures and Tables*

Figures and tables may be formatted horizontally (a.k.a. landscape) as long as their captions appear horizontal also. LaTeX will format landscape material for you if a couple of conditions are met. You have to have a printer and printer driver that allow rotations and you have to have a couple of add-on LaTeX packages.

Include the rotating package

```
\usepackage[figuresright]{rotating}
```

and read the documentation that comes with the package.

Figure 9.3 is an example of how a landscape table might be formatted.

```
\begin{figure}[p]
   \begin{leftfullpage}
      \caption{ . . . }
   \end{leftfullpage}
\end{figure}
\begin{figure}[p]% the right side space
  \begin{xtrafullpage}

     . . .

      ( note.. no caption here )
  \end{xtrafullpage}
\end{figure}
```

Figure 9.2b: This text would create a facing caption page with the accompaning figure in the one-side style.

```
\begin{sidewaystable}
   ...
   \caption{ . . . }
\end{sidewaystable}
```

Figure 9.3: This text would create a landscape table with caption.

9.2.5   Figure and Table Captions

Most captions are formatted with the \caption macro as described by Lamport[**?**, sec. C.9]. The uwthesis class extends this macro to allow continued figures and tables, and to provide multiple figures and tables with the same number, e.g., 3.1a, 3.1b, etc.

To format the caption for the first part of a figure or table that cannot fit onto a single page use the standard form:

```
\caption[toc]{text}
```

To format the caption for the subsequent parts of the figure or table use this caption:

```
\caption(-){(continued)}
```

It will keep the same number and the text of the caption will be *(continued)*.

To format the caption for the first part of a multi-part figure or table use the format:

```
\caption(a)[toc]{text}
```

The figure or table will be lettered (with 'a') as well as numbered. To format the caption for the subsequent parts of the multi-part figure or table use the format:

```
\caption(x){text}
```

where *x* is b, c, ... The parts will be lettered (with 'b', 'c', ...).

9.3   The Preliminary Pages

These are easy to format only because they are relatively invariant among theses. Therefore the difficulties have already been encountered and overcome by LaTeX and the thesis document classes.

Start with the definitions that describe your thesis. This sample thesis was printed with the parameters:

```
\Title{The Suitability of the \LaTeX\ Text Formatter\\
    for Thesis Preparation by Technical and\\
    Non-technical Degree Candidates}
\Author{Jim Fox}
\Program{UW Information Technology}
\Year{2012}

\Chair{Name of Chairperson}{title}{Chair's department}
\Signature{First committee member}
\Signature{Next committee member}
```

```
\Signature{etc}
```

Use two or more \Chair lines if you have co-chairs.

9.3.1   Copyright page

Print the copyright page with \copyrightpage.

9.3.2   Title page

Print the title page with \titlepage. The title page of this thesis was printed with[1]

```
\titlepage
```

You may change default text on the title page with these macros. You will have to redefine $\degree$text, for instance, if you're writing a Master's thesis instead of a dissertation.[2]

\Degree{*degree name*} defaults to "Doctor of Philosophy"

\School{*school name*} defaults to "University of Washington"

\Degreetext{*degree text*} defaults to "A dissertation submitted …"

\textofCommittee{*committee label*} defaults to "Reading Committee:"

\textofChair{*chair label*} defaults to "Chair of the Supervisory Committee:"

These definitions must appear <u>before</u> the \titlepage command.

---

[1] Actually, it wasn't. I added a footnote—something you would not do.

[2] If you use these they can be included with the other information before copyrightpage".

9.3.3 Abstract

Print the abstract with `\abstract`. It has one argument, which is the text of the abstract. All the names have already been defined. The abstract of this thesis was printed with

```
\abstract{This sample . . . 'real' dissertation.}
```

9.3.4 Tables of contents

Use the standard LaTeX commands to format these items.

9.3.5 Acknowledgments

Use the `\acknowledgments` macro to format the acknowledgments page. It has one argument, which is the text of the acknowledgment. The acknowledgments of this thesis was printed with

```
\acknowledgments{The author wishes . . . {\it il miglior fabbro}.\par}}
```

Chapter 10

RUNNING LATEX
(*AND PRINTING IF YOU MUST*)

TEX has been designed to produce exactly the same document on all computers and on all printers. *Exactly the same* means that the various spacings, line and page breaks, and even hyphenations will occur at the same places when the document is formatted on a variety of computers. However, the way you edit text files and run LATEX varies from system to system.

10.1 Running

Unfortunately, the author is woefully out of water where TEX on Windows is concerned. Google would be his resource. On a UNIX system he types

```
$ pdflatex uwthesis
```

and it generally works.

10.2 Printing

All implementations of TEX provide the option of **pdf** output, which is all the Graduate School now requires. Even if you intend to print a copy or two of your thesis—the best way to admire it—create a `pdf` anyway. It will print anywhere.

BIBLIOGRAPHY

[1] Keisha D. Carlson, Peter H. Sudmant, Maximilian O. Press, Evan E. Eichler, Jay Shendure, and Christine Queitsch. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Research*, 25(5):750–761, May 2015.

[2] Jennifer R Gatchel and Huda Y Zoghbi. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature reviews. Genetics*, 6(10):743–55, October 2005.

[3] Rita Gemayel, Marcelo D Vinces, Matthieu Legendre, and Kevin J Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, 44:445–77, January 2010.

[4] Harry T. Orr. Polyglutamine neurodegeneration: Expanded glutamines enhance native functions. *Current Opinion in Genetics and Development*, 22(3):251–255, 2012.

[5] Maximilian O. Press, Keisha D. Carlson, and Christine Queitsch. The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics*, 30(11):504–512, August 2014.

[6] Maximilian O Press, Hui Li, Nicole Creanza, Günter Kramer, Christine Queitsch, Victor Sourjik, and Elhanan Borenstein. Genome-scale co-evolutionary inference identifies functions and clients of bacterial Hsp90. *PLoS genetics*, 9(7):e1003631, 2013.

[7] P. Rival, M. O. Press, J. Bale, T. Grancharova, S. F. Undurraga, and C. Queitsch. The Conserved PFT1 Tandem Repeat is Crucial for Proper Flowering in Arabidopsis thaliana. *Genetics*, 198(2):747–754, August 2014.

[8] Soledad Francisca Undurraga, Maximilian Oliver Press, Matthieu Legendre, Nora Bujdoso, Jacob Bale, Hui Wang, Seth J Davis, Kevin J Verstrepen, and Christine Queitsch. Background-dependent effects of polyglutamine variation in the Arabidopsis thaliana gene ELF3. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47):19363–7, November 2012.

Appendix A

SUPPORTING CHAPTER 2

Appendix B

WHERE TO FIND THE FILES

The uwthesis class file, `uwthesis.cls`, contains the parameter settings, macro definitions, and other TEXnical commands which allow LaTeX to format a thesis. The source to the document you are reading, `uwthesis.tex`, contains many formatting examples which you may find useful. The bibliography database, `uwthesis.bib`, contains instructions to BibTeX to create and format the bibliography. You can find the latest of these files on:

- My page.

  `http://staff.washington.edu/fox/tex/uwthesis.html`

- CTAN

  `http://tug.ctan.org/tex-archive/macros/latex/contrib/uwthesis/`

  (not always as up-to-date as my site)